DOCUMENT RESUME

ED 185 090 TM 800 089

AUTHOR Haebara, Tomokazu
TITLE A Method for Investigating Item Bias Using Birnbaum's Three-Parameter Logistic Model. Iowa Testing Programs Occasional Papers Number 25.
INSTITUTION Iowa Testing Programs, Iowa City.
PUB DATE Dec 79
NOTE 55p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Achievement Tests; Black Students; Grade 9; *Item Analysis; Junior High Schools; *Statistical Significance; *Test Bias; Test Items; White Students
IDENTIFIERS Item Characteristic Curve Theory; Latent Trait Theory; *Three Parameter Model

ABSTRACT

It is proposed that statistical tests for item bias, when they are based on the three-parameter logistic item characteristic curve, should begin by testing for differences in the third parameter, the pseudo-chance level or "guessing" parameter. Other item-bias studies that have not considered the third parameter are discussed. Asymptotic significance tests for the three-item parameters are described. An experiment is described based on the 60-item Reading Comprehension Subtest of the Tests of Achievement and Proficiency, using the results for 400 black and 400 white ninth-grade students. None of the 60 items showed any indication of bias in the pseudo-chance level parameter so a weighted average parameter was computed and used in the remainder of the study. Comparisons of the difficulty and discrimination parameters did indicate some biased items. Some interpretation of the results are provided in relation to the content of the items. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED185090

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

IOWA TESTING PROGRAMS OCCASIONAL PAPERS

Number 25 - December 1979

A Method for Investigating Item Bias Using Birnbaum's Three-Parameter Logistic Model

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Tomokazu Haebara

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Tomokazu Haebara

Iowa Testing Programs

The University of Iowa

TM800089

ABSTRACT

Recently several methods for detecting item bias using Birnbaum's three-parameter logistic model have been proposed. But those methods are not satisfactory in that possible differences in the item pseudo-chance level parameters for different subpopulations are not taken into account. In this paper an approach to item bias identification that utilizes relevant information contained in the estimates of all three item parameters is proposed. This approach consists of a sequence of statistical tests of hypotheses of item parameter invariance based on the approximate sampling distributions of the maximum likelihood estimators of the parameters. The item responses of 400 black students and 400 white students to the Reading Comprehension Subtest of the Tests of Achievement and Proficiency (TAP) were analyzed by the proposed method. Some interpretations of the results are provided in relation to the content of the items.

TABLE OF CONTENTS

	Page
LIST OF TABLES	iii
LIST OF FIGURES	iv
CHAPTER	
I. PROBLEM	1
Purpose of This Study	8
II. METHOD	9
Three-Parameter Logistic Model	9
Lord's Method	11
Rudner's Method	12
Outline of Proposed Method	13
Asymptotic Significance Test	15
Equating	18
III. APPLICATION	23
Test and Subgroups Used	23
Significance Level	24
Pseudo-Chance Level Parameter	25
Discrimination and Difficulty Parameters	28
Interpretations of the Results	33
IV. DISCUSSION	41
BIBLIOGRAPHY	45

LIST OF TABLES

Table		Page
1.	Original and Equated Estimates of c_g and Critical Ratios	26
2.	Original and Equated Estimates of a_g and Critical Ratios	29
3.	Original and Equated Estimates of b_g and Critical Ratios	31

LIST OF FIGURES

Figure		Page
1.	ICC's for an Unbiased item (Item 28)	34
2.	ICC's with Significantly Different Estimates of b_g (Item 18)	35
3.	ICC's with Significantly Different Estimates of a_g (Item 55)	36
4.	ICC's with Significantly Different Estimates of a_g and b_g (Item 36)	38
5.	ICC's with Significantly Different Estimates of a_g and b_g (Item 51)	39

CHAPTER I

PROBLEM

In response to a growing concern about cultural, racial, and sex bias in, or in use of, educational tests, a substantial amount of investigation has been performed for detecting and reducing bias. Bias is such a multifaceted concept that various aspects of it have been emphasized by different researchers.

Green (1976) stated that biased tests are not necessarily bad because "test bias and unfairness are not the same thing. . . . Unfairness is a function of how a test is used while bias is a characteristic of the test" (p. 1). Although the present author recognizes the possibility that a wise user of a biased test may prevent its unfair use, he nevertheless supposes that complicated aspects of bias can be explicated most consistently in terms of how much and against whom a test or an item is unfair.

There are four main cases identified for investigating bias. First, there are situations in which an external criterion is available and bias is examined for a test as a whole (test bias). In this case, a test is considered biased if, among those who perform equally well on the criterion, the average performance on the test differs according to the subgroup membership (Cronbach, 1970, p. 305). This definition is usually intended and appropriate for use in intelligence or aptitude

tests, which are used for predicting some external criterion such as grade in school and success in an employment setting, or for selection purposes.

To examine test bias based on this definition, regression analyses have been utilized (e.g., Reschly & Sabers, 1979). The two principal concerns in those regression analyses are slope bias and intercept bias (Anastasi, 1976, pp. 191-197). To understand conceptually the relationship between regression and test bias, the following explanation may suffice. When the lines of regression of the criterion scores on the test scores are drawn for two subgroups, and if one line is higher than the other over the whole range of possible test scores, then the test is considered to be biased (unfair) against the subgroup corresponding to the higher line. If the two lines are crossed, we have a situation of differential bias; that is, the test is biased against those in a subgroup who have ability higher than the crossing point and those in the other subgroup who have ability lower than that point. Similar analyses and interpretations can also be made by use of expectancy tables.

In this definition, the criterion measure is required to be unbiased (in the sense of the third case described below). Otherwise we would encounter such a cynical claim as "Culturally biased tests . . . show good predictive validity for culturally biased criteria" (Jensen, 1976, p. 340).

Second, there are situations in which an external criterion is available and bias is examined at an individual item level (item bias).

In this case, a definition of item bias similar to that for test bias in the first case applies. And item bias can be examined by comparing the values of conditional item difficulty indices (proportion of correct answers, if the item is scored dichotomously) given a certain score of the criterion measure among the subgroups for various scores of the criterion measure. Or, alternatively, one could compare the conditional distributions of the criterion scores given an item score among subgroups. The way of interpretation of the results in terms of how much and against whom an item is unfair will be obvious. This type of item analysis will provide valuable information to reduce test bias as well as item bias, if valid criterion measures on sufficiently large samples are available.

Third, there are situations in which no external criterion is available and bias is examined for a test as a whole (test bias). Although Green (1976) stated that achievement tests should (preferably) be validated against outside criteria, at least at present most bias studies in achievement tests are performed without external criteria. Actually, since achievement tests are constructed for measuring students' present status in the development of certain knowledge and skills, not for predicting external outcomes, their fairness for various subgroups ought to be checked against content and construct validity evidence rather than predictive validity.

In this case, a test can be considered to be unbiased if all individuals having the same underlying ability level have the same true score regardless of their subgroup membership. Here the term true

score is used in the sense as defined in classical test theory. This definition has frequently been expressed in a somewhat more ambiguous form: A test is considered to be unbiased if the test measures the same psychological trait across subgroups (Lord, 1976). Although the former definition is more directly related to the notion of fairness, it is the latter one that has facilitated various types of investigation. For example, in his effort to prove the unbiasness of intelligence tests, Jensen (1976) used such statistical indices as correlation of raw scores with age, internal consistency reliability, and rank order of item difficulty computed for both black and white samples and compared each of these indices between the two racial groups to show that most intelligence tests are measuring the same trait across racial groups.

As mentioned earlier, however, in trying to solve bias problems in the absence of an external criterion, the greatest emphasis ought to be placed on the examination of content validity and the continuous accumulation of relevant evidence of construct validity, rather than any single-shot statistical analysis. And once the agreement has been obtained that a test is reasonably unbiased as a whole, the test constructors' efforts are turned to detecting biased items and to improve further the fairness of the test either by eliminating those items or by revising them. Those item bias studies are useful not only for improving the particular test in which the items are contained but also for obtaining general ideas on what factors are tending to make items biased.

Such item bias studies in the absence of an external criterion represent the fourth case for investigating bias. In this case, "A test item can be considered to be unbiased if all individuals having the same underlying ability level have an equal probability of correctly answering the item, regardless of their subgroup membership" (Pine, 1977, p. 38). This definition of item bias is intended for dichotomously scored items, but it can easily be adapted to deal with multichotomously scored items. It is crucial to note that this type of item bias investigation is based on the fact, or at least the assumption, that the whole test is reasonably unbiased. To illustrate this idea, consider a test which is completely unbiased containing n unbiased items. Then one could obtain an estimate of underlying ability level for each individual based on his performance on this set of items. If the (n + 1)st item were added to the test, whether or not this new item is biased would be judged by examining the relationship between the performance on this item and the estimated underlying ability level for various subgroups. The same thing could be done even if a small number of biased items were contained in the original set of items, since the underlying ability would be determined largely by the majority of items which were unbiased. However, it is obvious that such an item bias study would be misleading or even could make the test worse if the whole test itself were biased.

In Pine's definition, if the total score were used as an estimate of the underlying ability, regressions of item scores on the total scores could be used to examine item bias. In fact, Scheuneman (1975)

used this procedure for black and white samples. In this method, however, a conditional probability of correctly answering an item given a certain total score is estimated only from the item responses of those who obtained that particular total score. This implies instability of the probability estimates, especially when the number of items is large relative to the number of individuals. In order to utilize the performances of all individuals, some functional form of the item-test regression must be specified. Unfortunately, item-test regression curves have such a variety of different shapes that it is unlikely that they can be fitted by a simple family of mathematical functions (Lord & Novick, 1968, pp. 363-364).

To overcome this and other difficulties (Lord, 1976), latent trait theory can be adopted. For dichotomously scored items and unidimensional underlying abilities, the regression of item score on the underlying ability (latent trait) scale is called the item characteristic curve (ICC). (Hereafter the discussion will be restricted to the situations in which items are scored dichotomously and the latent space is unidimensional, among various latent trait models.) Usually a family of mathematical functions for ICC's is specified prior to estimating the exact forms of the curves. (An exception to this scheme is found in Lord, 1970.) The logistic cumulative distribution function (cdf) with one or two parameters or a modified form of the logistic cdf with three parameters is usually used for the mathematical form for ICC's.

If conditions for a particular ICC model (Birnbaum, 1968; Hambleton, Swaminathan, Cook, Eigner, & Gifford, 1978; Lord & Novick, 1968) are met, Pine's definition of unbiased item is equivalent to requiring that a test item have the same ICC for all subgroups. This is also equivalent to requiring that the parameters which describe the ICC be invariant from subgroup to subgroup after equating the latent trait scales by linear transformations (Pine, 1977, p. 38; Lord & Novick, 1968, pp. 379-381).

Within the framework of ICC theory, the interpretation of the ICC's for an item estimated for various subgroups in terms of the notion of unfairness seems easy. However, one point should be made clear here. The term latent trait may seem to suggest that thus derived ability levels were unbiased. But this is not necessarily true. Consider a test containing items that are all biased against girls. As a result, the true score is also biased. And, not surprisingly, so is the latent trait score, since the same response pattern of boys and girls necessarily results in the same latent trait score estimate if the performances of boys and girls are analyzed simultaneously, and approximately so if the scales are linearly equated after separate analyses. ICC theory has no mystic power to transform a biased test score into an unbiased latent trait score. Therefore, the fact, or at least the assumption, that the test as a whole is fairly unbiased remains a crucial requisite for item bias investigation based on ICC theory as well.

Purpose of This Study

In the present study, a method for investigating item bias in the absence of an external criterion (the fourth case described above) is proposed. The method utilizes Birnbaum's (1968) three-parameter logistic model, which involves the pseudo-chance parameter (guessing parameter) as well as difficulty and discrimination parameters. The choice of the model is based on the author's opinion that in item bias investigation guessing behaviors are an important aspect to be considered if the items are of selection type, which the present study is to deal with. The basic idea of the proposed method is to examine the item parameter invariance across subgroups based on the approximate asymptotic standard errors of the maximum likelihood (ML) estimates of the ICC parameters. In the next chapter, this method will be explained in contrast to similar methods previously proposed by several researchers (Laksana, 1979; Lord, 1976; Rudner, 1977). Then, the method will be applied to a standardized achievement test using two racial subgroup samples, and its utility and drawbacks will be discussed in the subsequent chapters.

CHAPTER II

METHOD

Three-Parameter Logistic Model

Birnbaum (1968) proposed the three-parameter logistic model in which the mathematical form of the ICC is given by

$$P_g(\theta) = c_g + \frac{1 - c_g}{1 + \exp\{-Da_g(\theta - b_g)\}}. \quad (1)$$

In this equation, θ denotes the unidimensional ability (or latent trait) parameter. The parameters a_g, b_g, and c_g are respectively item discrimination, item difficulty, and item pseudo-chance level parameters of item g. The symbol D denotes an arbitrary scaling factor. It has been customary to take D = 1.7, since this value of the scaling factor provides very close agreement between the logistic cdf and the standard normal ogive. Once the item parameters and the scaling factor are specified, Equation 1 provides the probability that an examinee with ability θ answers the item correctly.

Conceptually, the three-parameter logistic model is related to the normal ogive (and two-parameter logistic) model through a very schematized psychological hypothesis; that is, "this model assumes that if an examinee has ability θ, then the probability that he will know the correct answer is given by a normal ogive function . . . it further

assumes that if he does not know it he will guess, and, with probability c_g, will guess correctly" (Birnbaum, 1968, p. 404).

Among descriptive characteristics of the three-parameter logistic ICC are:

$$\lim_{\theta \to -\infty} P_g(\theta) = c_g \tag{2}$$

$$P_g(\theta | \theta = b_g) = \frac{1}{2} + \frac{c_g}{2} \tag{3}$$

$$\left(\frac{\partial P_g(\theta)}{\partial \theta} \right)_{\theta = b_g} = \frac{D}{4} a_g (1 - c_g) \tag{4}$$

$$\frac{\partial^2 P_g(\theta)}{\partial \theta^2} = 0, \text{ if } \theta = b_g$$

$$> 0, \text{ if } \theta < b_g$$

$$< 0, \text{ if } \theta > b_g \tag{5}$$

Equation 2 indicates that the lower asymptote of the curve is represented by the value of c_g. Equations 3, 4, and 5 indicate that the curve has the steepest slope at $\theta = b_g$, the slope being $\frac{D}{4} a_g (1 - c_g)$ and the probability of correctly answering the item at that level on the ability scale is $\frac{1}{2} + \frac{c_g}{2}$. Note that, unlike the two-parameter model, the largest slope is not proportional to the value of the discrimination parameter, a_g.

These descriptive characteristics of the three-parameter logistic ICC may reveal that the considerations of the pseudo-chance level parameter, c_g, are very important in examining the invariance of an ICC from subgroup to subgroup. However, the important role of the

parameter c_g has not been fully recognized by the researchers who have used this model in their studies on item bias. This point is now amplified by reviewing two methods previously proposed for item bias investigation.

Lord's Method

In his study on item bias in the Verbal Section of the Scholastic Aptitude Test using black and white samples, Lord (1976) first estimated, for each group separately, the parameters a_g, b_g, and c_g of each item and θ of each examinee simultaneously, using a computer program, LOGIST (Wood, Wingersky, & Lord, 1976). Next, he plotted the values of b_g for the black group against those for the white group for equating the scale. (The principle for equating will be explained in a later section.) As a third step, Lord performed an asymptotic significance test to test the null hypothesis that a particular item has the same values of a_g and b_g for blacks and whites. (Although he did not give a mathematical justification for the use of the asymptotic test, he showed an empirical support for his procedure.) Then, Lord proceeded to reestimate θ of each examinee for the total group containing both blacks and whites, using only those items that were not rejected by the asymptotic test. He then reestimated item parameters for each group separately, using the ability parameter values obtained in the previous step. Finally, Lord repeated the asymptotic test using those item parameter estimates to identify biased items.

The problem that concerns the present author lies in the fact that Lord completely ignored information contained in the estimates of c_g.

Lord simply maintained that the value of the parameter c_g was required to be the same for blacks and whites. However, if the true value of c_g for blacks were different from that for whites, the two ICC's would, of course, never be identical even if a_g and b_g were identical. And if, in spite of the true difference in c_g, a single value were assigned to both groups, then the statistical test of the null hypothesis that the two ICC's are identical would result in a spurious result.

Rudner's Method

Rudner (1977) compared item performances of the students in programs for the hearing impaired and those of normal hearing students on the Reading Comprehension Subtest of the Stanford Achievement Test. Rudner estimated item parameters for each group separately, using a minimum chi-square method. He then equated the item parameters by assigning the values of c_g for the normal hearing group to those for the hearing impaired group and by using regression procedures for equating a_g and b_g. Finally, Rudner computed the area between the ICC's for the two groups for each item and used it as an index of the amount of item bias.

Laksana (1979), in his study of racial and sex bias in the Iowa Tests of Basic Skills, computed an additional item bias index which is the area under an ICC for one group subtracted from the area under the corresponding ICC for the other group and called this index the signed aberrance. In Laksana's study, unlike Rudner's, the item parameters were estimated by an ML estimation procedure using LOGIST. To equate

c_g, he followed Rudner's procedure; for example, the values of c_g for blacks were replaced with those for whites for each item.

Rudner's and Laksana's methods are not satisfactory, since, again, they did not pay any attention to information contained in the estimates of c_g. The equating of c_g was performed regardless of the magnitude of difference in its estimates for two subgroups. And, unfortunately, they used the estimates for the higher performance group (the normal hearing group in Rudner's case and the white group in Laksana's case) to replace those for the lower performance group. It is intuitively obvious and, in certain cases, can be mathematically proven that, when two ICC's have the same true value of c_g, the parameter is estimated better by using the lower performance group, since it has the more examinees in the lower tail of the ability distribution than the higher performance group. In the case where the true value of c_g is invariant from subgroup to subgroup, an even better estimate could be obtained either by using the total combined group or by taking a weighted average of the estimates obtained from the two or more subgroups with the weights being the relative precision of each estimate.

Outline of Proposed Method

The steps constituting the proposed method for detecting biased items are outlined below. Some of the steps will be discussed in more detail in the subsequent sections. Hereafter it is assumed that there are only two subgroups involved in a study, although the same sequence

of procedures can easily be applied to the case in which more than two subgroups are involved.

(1) Using LOGIST, the parameters a_g, b_g, and c_g of each item and θ of each examinee are estimated for each subgroup separately.

(2) Approximate values of asymptotic standard errors of the estimates of c_g are computed.

(3) Asymptotic significance tests are performed for each item to see whether the values of c_g estimated from the two subgroups are significantly different. Those items which show a statistically significant difference between the two estimates of c_g are considered to be biased with respect to the pseudo-chance level of correctly answering the item.

(4) For each of those items that were not considered to be biased with respect to c_g in Step 3, a single value of c_g is assigned to the ICC's for both groups (equating of c_g). This can be done either (i) by reestimating c_g by LOGIST, using the total combined group of examinees and only those items that were not rejected in Step 3; (ii) by taking a weighted average of the estimates obtained from the two subgroups, weighted by their respective relative precision; or (iii) by using the estimates obtained from the lower performance subgroup.

(5) Using the values of c_g obtained in Step 4 as conditioning values, the parameters a_g and b_g are reestimated for each subgroup. If, in Step 4, the first procedure was employed for determining the common values of c_g, the ability parameters obtained in that process are also used as conditioning values.

(6) The parameters a_g and b_g estimated in Step 5 are equated.

(7) Steps 1, 2, and 3 are repeated for a_g and b_g. The rejected items are considered to be biased with respect to the item discrimination or with respect to the item difficulty.

(8) Final decisions on item bias are made on the basis of the results obtained above, visual inspections of the forms of ICC's, and examination of the format and content of items.

Asymptotic Significance Test

Hambleton et al. (1978, pp. 477-483) reviewed the findings on the consistency and other asymptotic properties of ML estimators for ICC parameters. In particular, it was shown that, if item parameters and ability parameters were simultaneously estimated in the three-parameter logistic model, asymptotic sampling distributions of those ML estimators could not be identified in exact forms.

However, it can be shown that, if only one item parameter (a_g, b_g, or c_g) of an item is unknown, the ML estimator for that parameter is asymptotically (i.e., as the number of examinees increases infinitely) distributed (from sample to sample) normally with mean equal to its true value. In this case the variance of the asymptotic distribution is expressed in terms of the known parameters and the true value of the parameter being estimated. Below are those asymptotic variances:

$$V(\hat{a}_g) = \left(\frac{D^2}{(1 - c_g)^2} \sum_a (\theta_a - b_g)^2 (P_{ga} - c_g)^2 \frac{1 - P_{ga}}{P_{ga}} \right)^{-1} \quad (6)$$

$$V(\hat{b}_g) = \left(\frac{D^2 a_g^2}{(1 - c_g)^2} \sum_a (P_{ga} - c_g)^2 \frac{1 - P_{ga}}{P_{ga}} \right)^{-1} \quad (7)$$

$$V(\hat{c}_g) = \left(\frac{1}{(1 - c_g)^2} \sum_a \frac{1 - P_{ga}}{P_{ga}} \right)^{-1} \quad (8)$$

Here, subscripts g and a refer to item and examinee, respectively. Each of these variances is derived by taking the expectation of the second derivative of the log-likelihood function differentiated with respect to the parameter being estimated and by changing the sign. The normality of the asymptotic sampling distributions under the restricted assumption as stated above can be proven by incorporating Liapounov's theorem (Hogg & Craig, 1978, p. 317) into the proof for the normality of the asymptotic distributions of ML estimators for identically distributed random variables (Cramér, 1946, pp. 500-506; Kendall & Stuart, 1973, pp. 45-46). The regularity conditions for Liapounov's theorem are satisfied in the three-parameter logistic model.

In the case of simultaneous estimation, these asymptotic variances might not hold, and the asymptotic distributions might deviate from the normal distributions. In spite of the lack of mathematical supports, however, the author decided to take the normal distributions with means equal to the true values of the parameters being estimated and variances equal to those shown in Equations 6, 7, and 8, as fairly good approximations for the true sampling distributions. This

decision is based on the empirical evidence that ML estimates of parameters in the three-parameter logistic model converge reasonably well to their true values (Lord, 1975b), and on the asymptotic properties of ML estimators of several unknown parameters simultaneously estimated in the case of identical random distributions; that is, if a random variable is identically distributed with several unknown parameters, "the likelihood equations have a system of solutions which are asymptotically normal and joint asymptotically efficient estimates of the parameters" (Cramér, 1946, p. 504). Throughout this study, the independence of the ML estimator of a parameter from the estimators of the other parameters is also assumed so that a significance test for the invariance of a parameter can be performed independently from tests for the other parameters, although the independence might not hold strictly. In summary, the asymptotic significance tests employed in the proposed method are based on approximations of the sampling distributions of ML estimators for ICC parameters and, therefore, should not be considered to be exact tests.

Keeping these points in mind, the approximate significance tests are performed first on the invariance of c_g for the two subgroups (Steps 2 and 3). From Equation 8, the approximate standard errors of the estimates of c_g are computed for each subgroup. (These standard errors are also utilized in LOGIST in an effort to avoid wild fluctuations of the estimates, Wood, et al., 1976, p. 17.) If the subgroup samples are random samples from the respective populations, the ordinary significance test of the hypothesis of equal means of two normal

distributions is adopted by setting an appropriate α level. The items rejected through this procedure are considered to be biased with respect to c_g. That is, they are identified as biased in the sense that one subgroup has a higher base line (lower asymptote) of the ICC than the other subgroup. The difference in c_g may be interpreted in terms of the difference in the probability of correctly answering the item when an examinee does not know the correct answer. The rejected items will not be tested on the invariance of a_g and b_g, since there is no way to do so when the true value of c_g is different from subgroup to subgroup. However, additional information on the characteristics of those items may be obtained by examining visually the forms of the ICC's.

Equating

In order to make comparisons of a_g and b_g across subgroups possible, equating of c_g and the ability scale is necessary. The equating of c_g can be done by using one of the three procedures listed in Step 4. It can be seen in Table 1 in Chapter III that the first and the second procedures yield very similar values for the common (equated) values of c_g. An advantage of the first procedure (estimating c_g from the total combined sample) is that the estimates of θ obtained in that estimation process can be used as conditioning values when the values of a_g and b_g are reestimated for each subgroup in Step 5. The values of a_g and b_g obtained in this way are already expressed on the common ability scale. Thus Steps 5 and 6 are performed at the same time. An advantage of the second procedure (weighted average) is

that a run of the LOGIST program for the total group can be bypassed; the cost of the run may be very expensive when the sample size is large.

In the second procedure, the weights are the relative precision of the estimates of c_g obtained from the two subgroups separately. With the parameter and precision (p) for one subgroup being designated by an asterisk, the relative precision for each subgroup is determined by the following formulas:

$$p = \frac{V(\hat{c}_g)^{-1}}{V(\hat{c}_g)^{-1} + V(\hat{c}_g^*)^{-1}} \tag{9}$$

$$p^* = \frac{V(\hat{c}_g^*)^{-1}}{V(\hat{c}_g)^{-1} + V(\hat{c}_g^*)^{-1}} \tag{10}$$

This procedure to take a weighted average is analogous to a Bayesian procedure, in which the mean of a posterior distribution for a normal mean, in the case of known variance, is determined by taking a weighted average of the prior mean and the sample mean with weights being the respective precision (Novick & Jackson, 1974, p. 140).

An obvious advantage of the third procedure (using the estimates from the lower performance subgroup) is that the reestimation of a_g and b_g for the lower performance subgroup can be bypassed when no items were rejected in Step 3. However, this procedure is permitted only if the estimates from the lower performance subgroup are very close to the values that would be obtained by using the first or the

second procedure. This condition is equivalent to requiring that the two subgroups under study be substantially different regarding the average performances on the test so that the relative precision for the lower performance subgroup is close to unity. (The reason why the lower performance subgroup yields the more precise estimates of c_g is clear in view of Equation 8.)

The equating of a_g and b_g (Steps 5 and 6) is performed by using only those items that were not rejected in Step 3. The equating procedure employed when the first procedure to determine the common c_g was used in Step 4 was already discussed above. If the second or the third procedure was employed in Step 4, one must first reestimate a_g and b_g for each subgroup with the values of c_g used as conditioning values. (As mentioned above, the reestimation for the lower performance subgroup is omitted if the third procedure was employed.) Then one proceeds to equate a_g and b_g obtained in this way.

A property of the model is that the ability scale derived has an arbitrary origin and an arbitrary unit. With the parameters for one subgroup being designated by an asterisk, the estimated values of θ and θ^* in the two subgroups are equated so that the equation

$$P_g(\theta) = P_g^*(\theta^*) \tag{11}$$

holds for all items. Now that we have

$$c_g = c_g^*, \tag{12}$$

Equation 11 is equivalent to

$$a_g(\theta - b_g) = a_g^*(\theta^* - b_g^*) \tag{13}$$

for all items (See Equation 1). From this we have

$$\theta = k\theta^* + d, \tag{14}$$

$$a_g = \frac{1}{k}a_g^*, \tag{15}$$

and

$$b_g = kb_g^* + d, \tag{16}$$

where

$$k = \frac{a_g^*}{a_g}, \tag{17}$$

and

$$d = b_g - kb_g^*. \tag{18}$$

It is clear from Equation 14 that the values of k and d should be invariant from item to item. There have been various procedures employed to determine these values (Lord, 1976; Rudner, 1977; Shiba, 1978). The simplest way is to utilize means and standard deviations of estimated b_g and b_g^*, where

$$k = \frac{\text{SD of } b_g}{\text{SD of } b_g^*}, \tag{19}$$

and

$$d = \bar{b}_g - k\bar{b}_g^*. \tag{20}$$

This procedure is perhaps better than that which estimates k from the values of a_g and a_g^*, since the difficulty parameter is typically better estimated than the discrimination parameter.

One might argue that we could equate a_g and b_g by using originally estimated values of these parameters without reestimating them

with equated values of c_g used as the conditioning values. However, the original a_g and b_g are not solutions of any likelihood equation if the simultaneously estimated c_g is changed to a different value. Although this problem is not very serious when the difference between the original c_g and the equated c_g is small, the author prefers to find the solutions of the conditional likelihood equations, where the condition is the use of the equated values of c_g as known parameters. Tables 2 and 3 in Chapter III will show that the differences between the parameters reestimated in this way and their original values are not negligible.

Finally, significance tests on a_g and b_g are performed to identify items that are biased with respect to item discrimination and those that are biased with respect to item difficulty. Comprehensive evaluation on item bias is made on the basis of the results of the significance tests, visual inspections of ICC's, and careful review of items.

CHAPTER III

APPLICATION

Test and Subgroups Used

In this chapter, the method proposed is applied to the 1978 national standardization data on the Reading Comprehension Subtest of the Tests of Achievement and Proficiency (TAP), Form 7, Level 15. The test (hereafter the Reading Comprehension Subtest is simply referred to as the test) consists of eight reading selections with a total of 60 five-alternative multiple-choice items.

From the data set, two subgroups were constructed on the basis of the examinees' racial membership. One subgroup, the black group, consists of 400 black ninth graders with 200 males and 200 females. The other group, the white group, consists of 400 white ninth graders with 200 males and 200 females. In order to avoid introducing irrelevant differences between the two groups such as locations and sizes of the schools represented, both groups were constructed by drawing equal numbers of blacks and whites (and males and females) from each of the schools, which are located in various parts of the country.

The computer program LOGIST was run for each group separately to estimate item parameters. The estimates will be shown in later sections (Tables 1, 2, and 3) with items grouped by the reading selections to which they belong.

Significance Level

The problems encompassing statistical significance, practical significance, and statistical power should be considered in determining the significance level (α level). To specify practically significant differences in parameter values for two groups, one must set lower limits of intolerable differences. For each of a_g, b_g, and c_g, one can estimate powers of the statistical test associated with various α levels by using the lower limits of intolerable differences as alternative hypotheses and by using the approximate standard errors of differences introduced earlier as fairly good approximations to the standard errors under the alternative hypotheses.

However, the consideration of power is accompanied by some difficulty. That is, if one chooses such α levels that a specified magnitude of power, say, 0.80, is assured, then one should come up with different α levels from item to item, reflecting the different values of standard errors for the items. Actually, the standard errors of differences vary substantially from item to item: They vary from 0.0964 to 1.3729 for a_g, from 0.0831 to 0.3724 for b_g, and from 0.0265 to 0.1058 for c_g.

To avoid this difficulty and to make simple the illustration of the proposed method, the conventional significance levels, five percent and one percent, are used throughout the present study. By doing this, we have the average critical difference of 0.49 for a_g, 0.33 for b_g, and 0.11 for c_g under the five percent level of significance; and the average critical difference of 0.64 for a_g, 0.43 for b_g, and 0.14 for

c_g under the one percent level of significance. Of course, some items are declared to be significant with smaller differences than the average and other items require larger differences to be significant, depending on the associated standard errors of differences. Certainly, the use of the same α level across all items is not a very satisfactory solution to the complicated problem. The complexity is left as an open question and will not be discussed in any more detail in this paper.

Pseudo-Chance Level Parameter

The first two columns of Table 1 show the estimates of c_g for the black and the white groups. With a few exceptions (Items 33, 54, and 60), the estimates are higher for the white group than for the black group. In view of the fact that, on the average, the white group performed better on the test than the black group with the respective mean raw scores being 34.8 and 23.8, it seems that the better performance group has the higher values of c_g. However, for any of the items, the difference in the estimates of c_g between the two groups was not significant at the five percent level. This is seen by looking at the third column of Table 1, which shows the critical ratios (or z values) for all items, each of which was computed by subtracting c_g for blacks from c_g for whites and by dividing the difference by the standard error of difference.

Since no significant difference was found for any item, an equated value of c_g for each and every item was sought. The fourth column of Table 1 shows the estimates of c_g obtained from the total combined

Table 1

Original and Equated Estimates of c_g and Critical Ratios

Item	Blacks	Whites	Critical Ratio	From Total Sample	Weighted Average
Selection 1					
1	0.1500	0.1750	0.25	0.1550	0.1555
2	0.1500	0.1750	0.63	0.1550	0.1575
3	0.1500	0.1750	0.41	0.1550	0.1585
4	0.1500	0.1750	0.48	0.1550	0.1593
5	0.1000	0.1750	1.44	0.1550	0.1188
6	0.1500	0.1750	0.41	0.1550	0.1565
Selection 2					
7	0.1500	0.1750	0.33	0.1550	0.1568
8	0.1500	0.1750	0.24	0.1550	0.1543
9	0.1500	0.1750	0.41	0.1550	0.1578
10	0.1500	0.1750	0.32	0.1550	0.1580
11	0.1500	0.1750	0.57	0.1550	0.1570
Selection 3					
12	0.1500	0.1750	0.36	0.1550	0.1588
13	0.1500	0.1750	0.44	0.1550	0.1563
14	0.1500	0.1750	0.53	0.1550	0.1580
15	0.1500	0.1750	0.37	0.1550	0.1595
16	0.1500	0.1750	0.47	0.1550	0.1598
17	0.1500	0.1750	0.34	0.1550	0.1568
18	0.1500	0.1750	0.40	0.1550	0.1588
19	0.1500	0.1750	0.53	0.1500	0.1593
20	0.1500	0.1750	0.57	0.1550	0.1618
21	0.1500	0.1750	0.49	0.1550	0.1598
22	0.1500	0.1750	0.47	0.1100	0.1588
Selection 4					
23	0.1500	0.1750	0.52	0.1550	0.1578
24	0.1500	0.1750	0.50	0.1550	0.1568
25	0.1500	0.1750	0.56	0.1856	0.1578
26	0.1500	0.1750	0.40	0.1550	0.1568
27	0.1500	0.1750	0.48	0.1550	0.1620
28	0.1500	0.1750	0.52	0.1550	0.1578
29	0.1000	0.1500	1.34	0.1100	0.1160
30	0.1500	0.1750	0.65	0.1306	0.1570
31	0.1500	0.1750	0.56	0.1550	0.1585

Table 1 (Continued)

Item	Blacks	Whites	Critical Ratio	From Total Sample	Weighted Average
Selection 5					
32	0.1500	0.1750	0.36	0.1550	0.1575
33	0.1500	0.1500	0.00	0.1550	0.1500
34	0.1500	0.1750	0.56	0.1550	0.1600
35	0.1000	0.1500	1.45	0.1550	0.1220
36	0.1500	0.1750	0.42	0.1500	0.1595
37	0.1500	0.1750	0.59	0.2100	0.1598
38	0.1500	0.1750	0.43	0.1550	0.1580
39	0.1500	0.1750	0.72	0.1550	0.1603
40	0.1500	0.1750	0.55	0.1550	0.1598
41	0.1500	0.1750	0.45	0.1550	0.1590
42	0.1500	0.1750	0.44	0.1550	0.1593
Selection 6					
43	0.1500	0.1750	0.53	0.1550	0.1568
44	0.1500	0.1750	0.39	0.1550	0.1575
45	0.1500	0.1750	0.52	0.1550	0.1588
46	0.1500	0.1750	0.30	0.1550	0.1573
47	0.1085	0.1500	1.57	0.1138	0.1263
Selection 7					
48	0.1500	0.1750	0.51	0.1550	0.1585
49	0.1500	0.1750	0.37	0.1550	0.1585
50	0.1500	0.1750	0.63	0.1550	0.1598
51	0.1500	0.1750	0.50	0.1500	0.1623
52	0.1500	0.1750	0.46	0.1550	0.1570
53	0.1500	0.1500	0.00	0.1750	0.1500
54	0.1785	0.1750	-0.11	0.1894	0.1769
55	0.1500	0.1750	0.44	0.1550	0.1605
Selection 8					
56	0.1500	0.1750	0.55	0.1550	0.1595
57	0.1500	0.1750	0.46	0.1550	0.1608
58	0.1500	0.1750	0.40	0.1550	0.1620
59	0.1500	0.1750	0.49	0.1550	0.1578
60	0.1915	0.1750	-0.43	0.1550	0.1847

group. The last column of the table contains the estimates obtained by taking the weighted averages of c_g for blacks and those for whites with the weights being the relative precision of the estimates. It is seen from the table that these two procedures yielded very similar values. Furthermore, these equated values appear to be so close to the estimates for the black group that one can justify the use of the original estimates for blacks as the common (equated) values of c_g. Therefore, it was decided to take the estimates for blacks as the equated values and to replace the estimates for whites with these values. By doing this, one can save computer time that would otherwise be needed to re-estimate parameters a_g and b_g for the black group.

Discrimination and Difficulty Parameters

The original estimates of a_g and b_g are shown in the first two columns of Tables 2 and 3, respectively. The third columns of the tables contain the values of a_g and b_g for the white group reestimated by using the estimates of c_g for the black group as conditioning values. The reestimated values for whites were then equated to the scale for blacks. The equated values for the white group are shown in the fourth columns of Tables 2 and 3. The last columns of the tables show the critical ratios computed by subtracting the original estimates for blacks from the equated estimates for whites and by dividing the differences by the standard errors of differences. Thus the positive values of the critical ratios indicate that the estimates for whites are greater than those for blacks. Significantly different estimates were

found in fifteen items for a_g and in seventeen items for b_g at the five percent level.

Interpretations of the Results

When none of the three item parameters, a_g, b_g, and c_g have significantly different estimates for the two groups, the two ICC's are almost identical, as shown in Figure 1. (Since the computer program plots a graph for whites and then plots a graph for blacks over the first graph, the lack of plots for whites for some ranges of θ indicates that the two groups have approximately identical values of $P_g(\theta)$ for those ranges.) Such an item can be considered unbiased if the assumptions stated earlier are satisfied.

Figure 2 shows the ICC's for an item which has significantly different estimates of b_g and approximately equal estimates of a_g for the two groups. For such an item, one group has higher probabilities of getting a correct answer than the other group for the whole range of θ. In this case, the item is judged as biased against the group with a higher value of b_g (the white group in this particular example). If, conversely, an item has significantly different estimates of a_g and approximately equal estimates of b_g, one group has higher probabilities of getting a correct answer for the values of θ above a certain point in the neighborhood of $\theta = b_g$ and lower probabilities for the values of θ below that point (Figure 3). Such an item can be called a differentially biased item.

Those items for which both a_g and b_g are significantly different show various patterns of discrepancies between the ICC's depending upon

Table 2

Original and Equated Estimates of a_g and Critical Ratios

Item	Blacks	Whites	Whites (Reesti-mated)	Whites (Equated)	Critical Ratio
Selection 1					
1	0.5790	0.5648	0.5960	0.7163	1.30
2	1.1572	0.8485	0.7863	0.9450	-0.82
3	0.5491	0.2762	0.2791	0.3354	-1.66
4	0.4846	0.7972	0.7406	0.8901	2.69**
5	1.1011	0.6045	0.5212	0.6264	-2.58**
6	0.5595	0.5806	0.5803	0.6974	1.05
Selection 2					
7	1.3783	0.9921	1.0162	1.2213	-0.77
8	1.2299	0.6214	0.6668	0.8014	-2.74**
9	0.8343	0.8020	0.7945	0.9548	0.72
10	0.5677	0.5127	0.5284	0.6350	0.61
11	2.1796	1.2200	1.1492	1.3811	-1.90
Selection 3					
12	0.8346	0.9252	0.9358	1.1246	1.79
13	0.7572	0.6252	0.6078	0.7305	-0.17
14	0.8783	0.7216	0.7026	0.8444	-0.17
15	1.0296	1.2500	1.2453	1.4966	2.30*
16	0.5761	0.7369	0.6987	0.8397	1.63
17	1.0880	0.6165	0.6229	0.7486	-2.19*
18	1.1330	0.8955	0.9002	1.0819	-0.27
19	0.7824	0.9502	0.9128	1.0970	1.54
20	0.4076	0.5929	0.5652	0.6793	2.04*
21	0.5259	0.5526	0.5256	0.6317	0.72
22	0.9461	1.2105	1.1857	1.4250	2.60**
Selection 4					
23	1.1023	1.0261	0.9848	1.1835	0.34
24	1.3401	0.7333	0.6721	0.8077	-2.11*
25	1.0061	0.9000	0.8549	1.0274	0.09
26	1.0228	0.9134	0.9043	1.0868	0.34
27	0.4455	0.5515	0.5464	0.6567	1.42
28	1.5054	1.2389	1.1975	1.4392	-0.22
29	1.0124	1.1503	1.0384	1.2479	1.02
30	1.9047	1.0418	0.9842	1.1828	-1.57
31	1.3822	1.7816	1.6498	1.9827	1.78

*$p < .05$
**$p < .01$

Table 2 (Continued)

Item	Blacks	Whites	Whites (Reesti-mated)	Whites (Equated)	Critical Ratio
Selection 5					
32	0.8098	0.8397	0.8205	0.9861	1.11
33	1.7072	1.1204	1.1134	1.3381	-0.86
34	0.6053	0.8832	0.8455	1.0161	2.21*
35	0.5407	0.4993	0.4246	0.5103	-0.32
36	1.3903	2.0721	1.9614	2.3572	2.86**
37	1.5582	2.3041	2.1348	2.5656	2.34*
38	0.7551	0.8608	0.8466	1.0174	1.48
39	1.9624	1.7583	1.6131	1.9386	-0.05
40	0.8469	0.7974	0.7730	0.9290	0.56
41	0.5427	0.5949	0.5659	0.6801	0.93
42	0.5161	0.6425	0.6226	0.7482	1.59
Selection 6					
43	1.5063	0.7840	0.7542	0.9064	-1.93
44	0.6917	0.7474	0.7519	0.9036	1.31
45	0.7994	0.6868	0.6376	0.7663	-0.17
46	1.0141	1.0564	1.0632	1.2778	1.44
47	1.3771	1.0771	0.7313	0.8789	-1.30
Selection 7					
48	1.1574	0.7358	0.6969	0.8375	-1.31
49	0.8395	0.9734	0.9405	1.1303	1.63
50	0.7397	0.6165	0.5650	0.6790	-0.35
51	0.8169	1.6080	1.5116	1.8166	3.52**
52	1.3792	0.9306	0.8911	1.0709	-1.12
53	1.7304	1.0306	1.0075	1.2108	-1.00
54	1.0421	1.4858	1.5276	1.8359	1.55
55	0.4536	0.9204	0.8604	1.0340	3.44**
Selection 8					
56	0.9775	0.7604	0.7328	0.8807	-0.40
57	0.7420	0.9045	0.8688	1.0441	1.45
58	0.3781	0.9449	0.8744	1.0509	4.04**
59	0.8939	0.4743	0.4461	0.5361	-1.68
60	4.0000	0.7876	0.8124	0.9763	-2.20*

*p < .05
**p < .01

Table 3

Original and Equated Estimates of b_g and Critical Ratios

Item	Blacks	Whites	Whites (Reesti-mated)	Whites (Equated)	Critical Ratio
Selection 1					
1	-0.4666	-2.1572	-2.0556	-0.8738	-2.10*
2	1.5173	0.4626	0.4136	1.1809	-2.19*
3	0.4823	-0.2542	-0.3524	0.5435	0.22
4	1.0957	-0.1011	-0.1695	0.6957	-2.01*
5	0.5732	-0.2604	-0.4981	0.4222	-1.07
6	0.8938	-0.7457	-0.7815	0.1864	-3.73**
Selection 2					
7	-0.3447	-1.2487	-1.2356	-0.1914	1.35
8	-0.6286	-2.3385	-2.1961	-0.9907	-2.24*
9	0.1928	-0.6808	-0.7162	0.2407	0.36
10	-0.3666	-1.3598	-1.3409	-0.2791	0.46
11	0.8030	-0.0173	-0.0560	0.7901	-0.15
Selection 3					
12	-0.2790	-0.8825	-0.8934	0.0933	2.84**
13	0.9350	-0.5749	-0.6343	0.3089	-3.82**
14	0.9786	0.0947	0.0465	0.8754	-0.70
15	-0.3943	-0.8463	-0.8622	0.1193	4.56**
16	0.6841	-0.0565	-0.1160	0.7402	0.33
17	-0.2725	-1.2178	-1.2355	-0.1919	0.57
18	-0.1482	-0.6662	-0.6859	0.2630	3.59**
19	0.9463	0.1204	0.0776	0.9013	-0.32
20	1.5699	0.8945	0.8363	1.5326	-0.15
21	0.9063	0.2020	0.1340	0.9482	0.21
22	0.4335	-0.3225	-0.3556	0.5408	0.70
Selection 4					
23	0.7610	-0.1511	-0.1945	0.6749	-0.74
24	0.7973	-0.1324	-0.2087	0.6630	-1.07
25	1.1316	0.1042	0.0553	0.8827	-1.83
26	0.1902	-0.8213	-0.8506	0.1289	-0.51
27	0.6358	0.4352	0.3726	1.1467	2.40*
28	0.6268	-0.1788	-0.2146	0.6581	0.33
29	1.3348	0.4207	0.3440	1.1229	-1.64
30	1.4261	0.3687	0.3263	1.1082	-2.64**
31	0.7531	-0.0708	-0.1122	0.7433	-0.11

*$p < .05$
**$p < .01$

Table 3 (Continued)

Item	Blacks	Whites	Whites (Reesti-mated)	Whites (Equated)	Critical Ratio
Selection 5					
32	0.0367	-0.9605	-1.0034	0.0018	-0.25
33	1.4849	0.8935	0.9044	1.5893	0.39
34	1.3155	0.3284	0.2843	1.0733	-0.78
35	2.0635	2.0709	1.9750	2.4801	1.53
36	-0.2213	-0.6983	-0.7311	0.2284	4.94**
37	0.6513	0.0550	0.0261	0.8584	2.49*
38	0.5357	-0.4875	-0.5266	0.4018	-0.93
39	1.1645	0.7526	0.7301	1.4442	2.88**
40	0.9519	0.3591	0.3139	1.0979	0.98
41	0.8300	-0.1497	-0.2203	0.6534	-0.90
42	0.6906	-0.2755	-0.3333	0.5594	-0.68
Selection 6					
43	1.0353	0.0541	0.0047	0.8406	-1.57
44	0.4383	-0.7017	-0.7229	0.2352	-1.28
45	1.0355	0.2301	0.1670	0.9757	-0.36
46	-0.5260	-1.4119	-1.4111	-0.3375	-2.13*
47	2.2317	2.1565	2.3306	2.7760	1.95
Selection 7					
48	0.7675	0.1663	0.1115	0.9295	1.19
49	-0.0044	-0.7939	-0.8398	0.1379	1.02
50	1.8085	1.0753	1.0342	1.6973	-0.48
51	0.3895	0.0395	0.0051	0.8409	3.46**
52	0.6061	-0.3297	-0.3811	0.5196	-0.73
53	1.6895	1.2240	1.2439	1.8717	1.11
54	2.5883	1.3039	1.3215	1.9363	-1.75
55	0.8430	-0.1925	-0.2515	0.6274	0.98
Selection 8					
56	1.0475	0.4500	0.4028	1.1719	0.77
57	0.5291	0.0129	-0.0313	0.8107	1.73
58	0.4973	-0.2486	-0.3166	0.5733	0.30
59	1.3528	0.3880	0.3135	1.0976	-1.18
60	1.2845	1.1081	1.1605	1.8024	3.74**

*p < .05
**p < .01

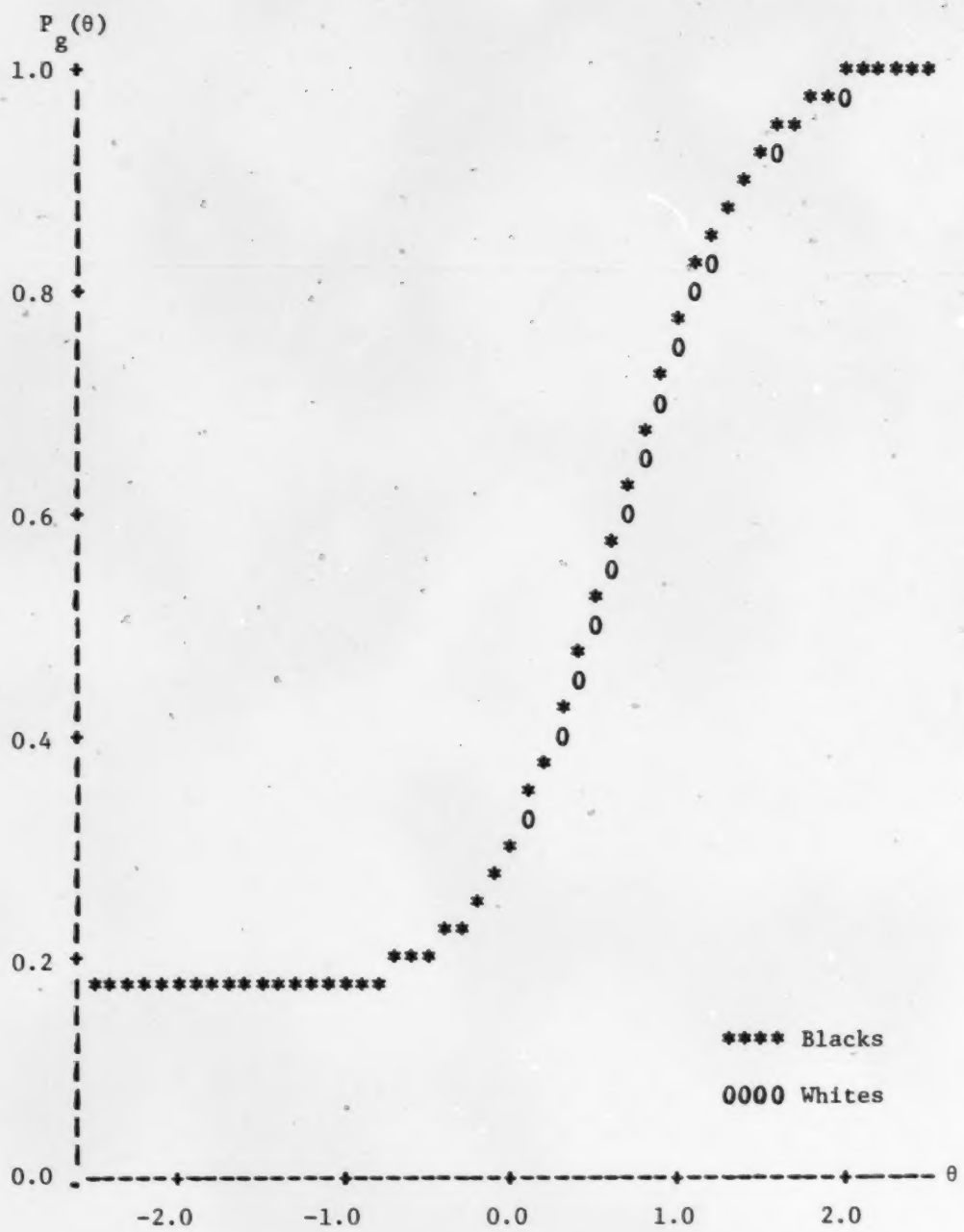

Figure 1. ICC's for an Unbiased Item (Item 28).

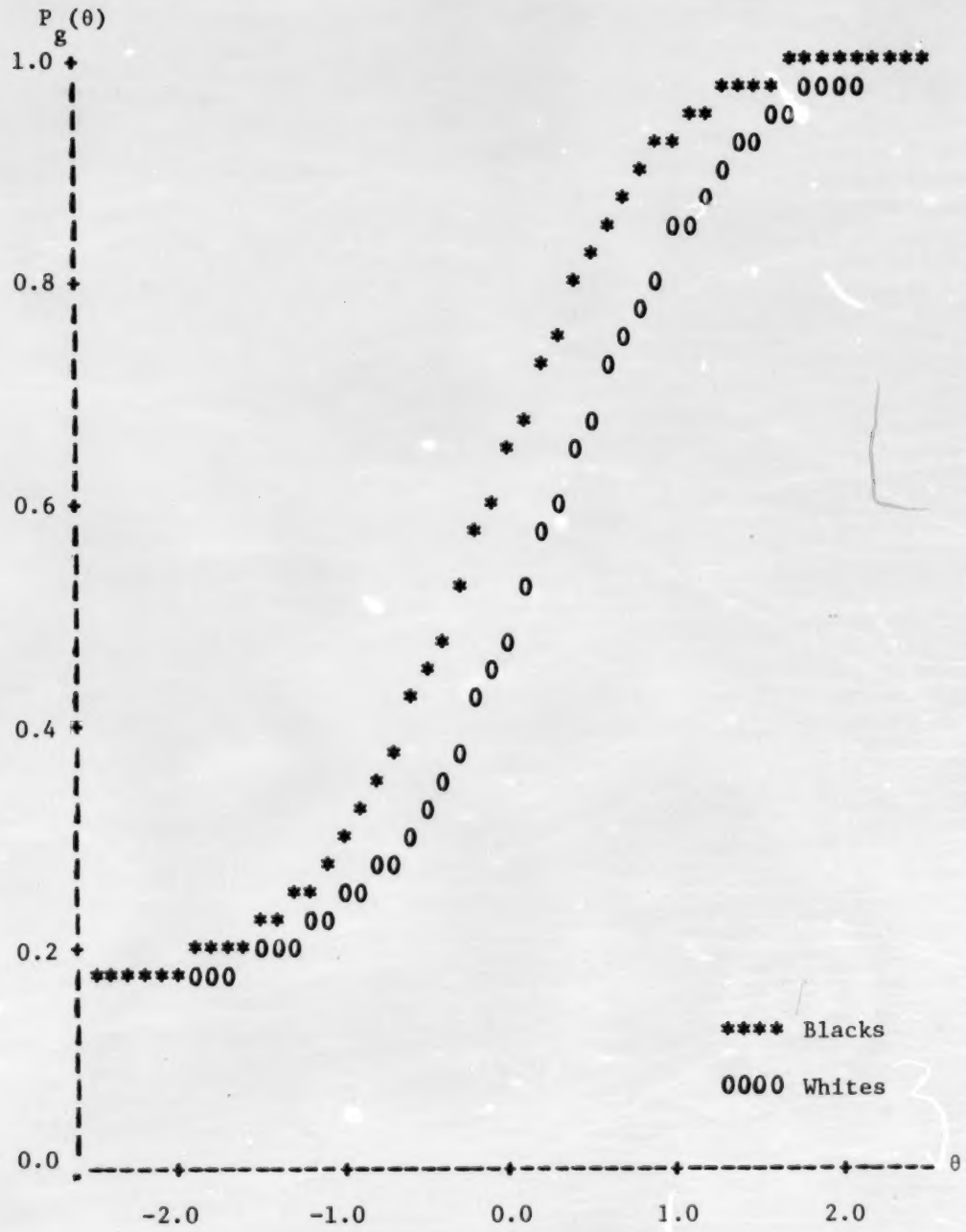

Figure 2. ICC's with Significantly Different Estimates of b_g (Item 18).

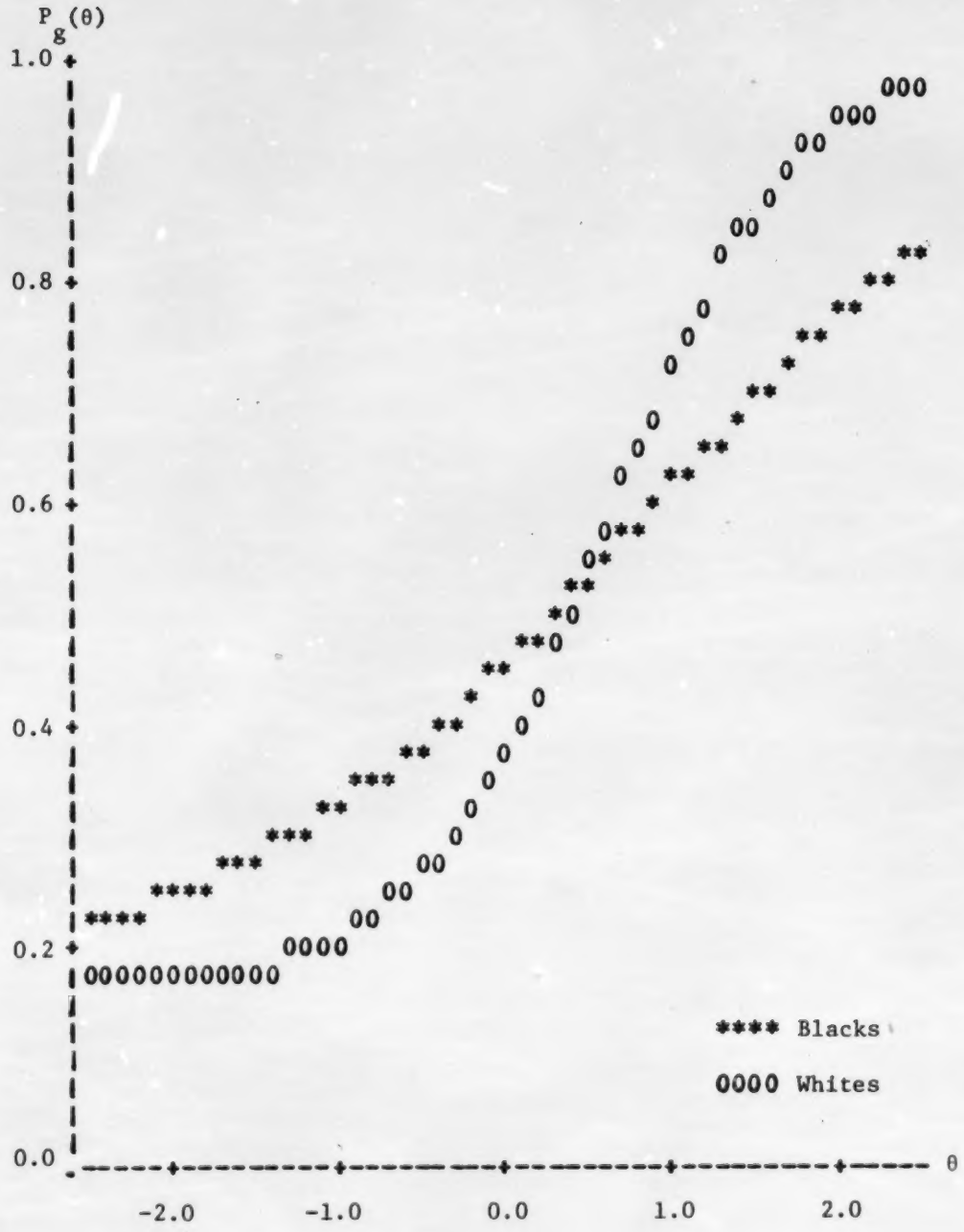

Figure 3. ICC's with Significantly Different Estimates of a_g (Item 55).

the values of a_g and b_g (for example, see Figures 4 and 5). In such a case, visual inspections of the ICC's in conjunction with consideration of the results from the significance tests will be necessary to judge on item bias.

As for the relationship between the results of the significance tests and the topics of the reading selections, some interesting findings were obtained. It is seen from Table 3 that four out of six items in Selection 1 have significantly higher values of b_g for blacks than for whites. Three of these four items have no significant differences in a_g (Table 2). This implies bias of items in the selection (or bias of the selection) against blacks. This selection is an advertisement to collect members of a club who are expected to purchase such goods as stereo records, 8-track tapes, and cassettes. The statistically significant results seem to be attributable to the content of the selection being biased against low SES students.

On the contrary, Selection 5 contains three items that have significantly higher values of b_g for whites and the graphs for those items imply bias against whites (the graph for one of the items is shown in Figure 4). This selection is a story about a black female banker and begins with a sentence which reads, "Frances was dark, like her father, and quite as tall."

If unbiasness of the whole test is assured, it can be said from the results that the identification of the races of characters in a story introduces a nuisance factor which is irrelevant to the objectives for which the story (reading selection) is presented and causes

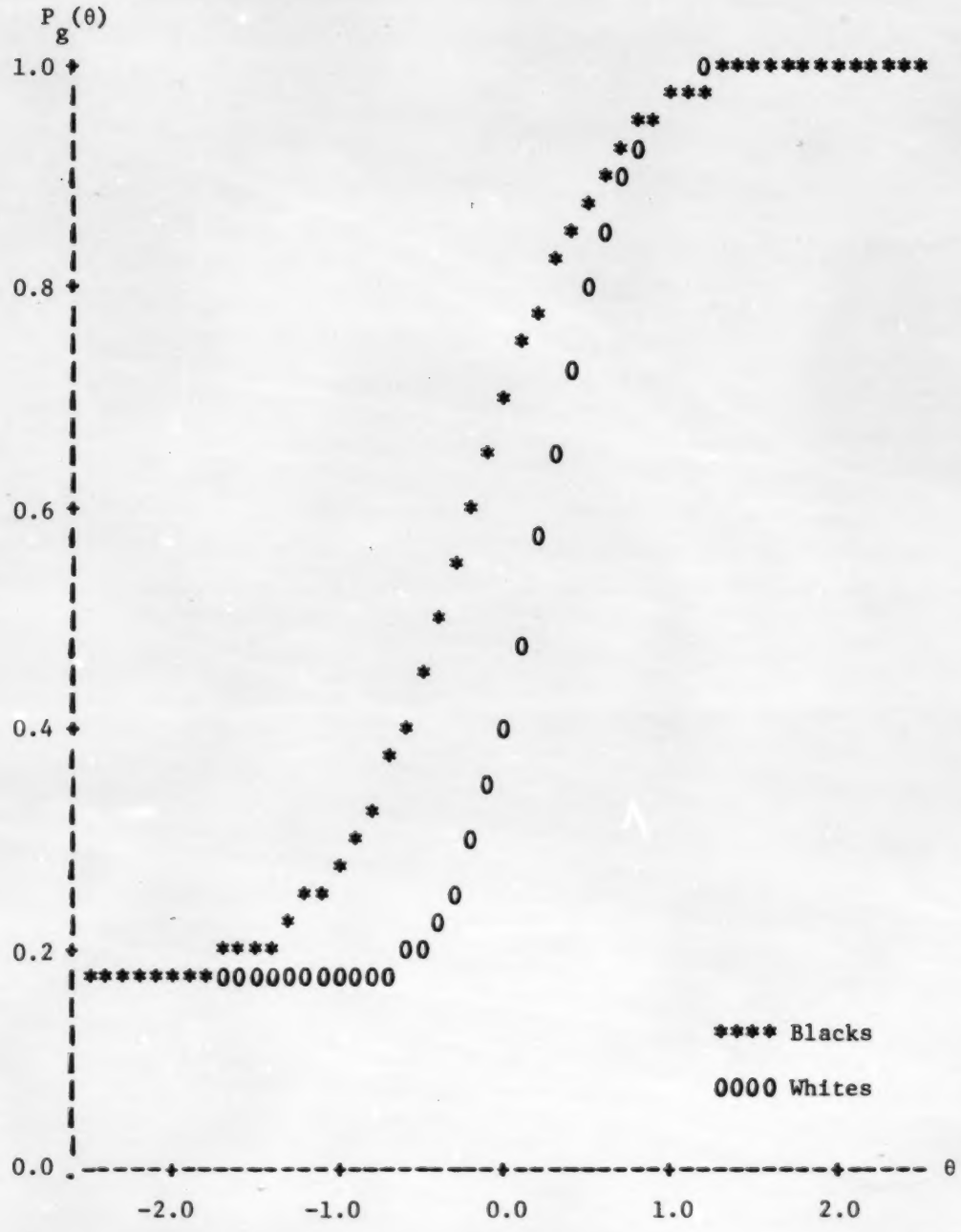

Figure 4. ICC's with Significantly Different Estimates of a_g and b_g (Item 36).

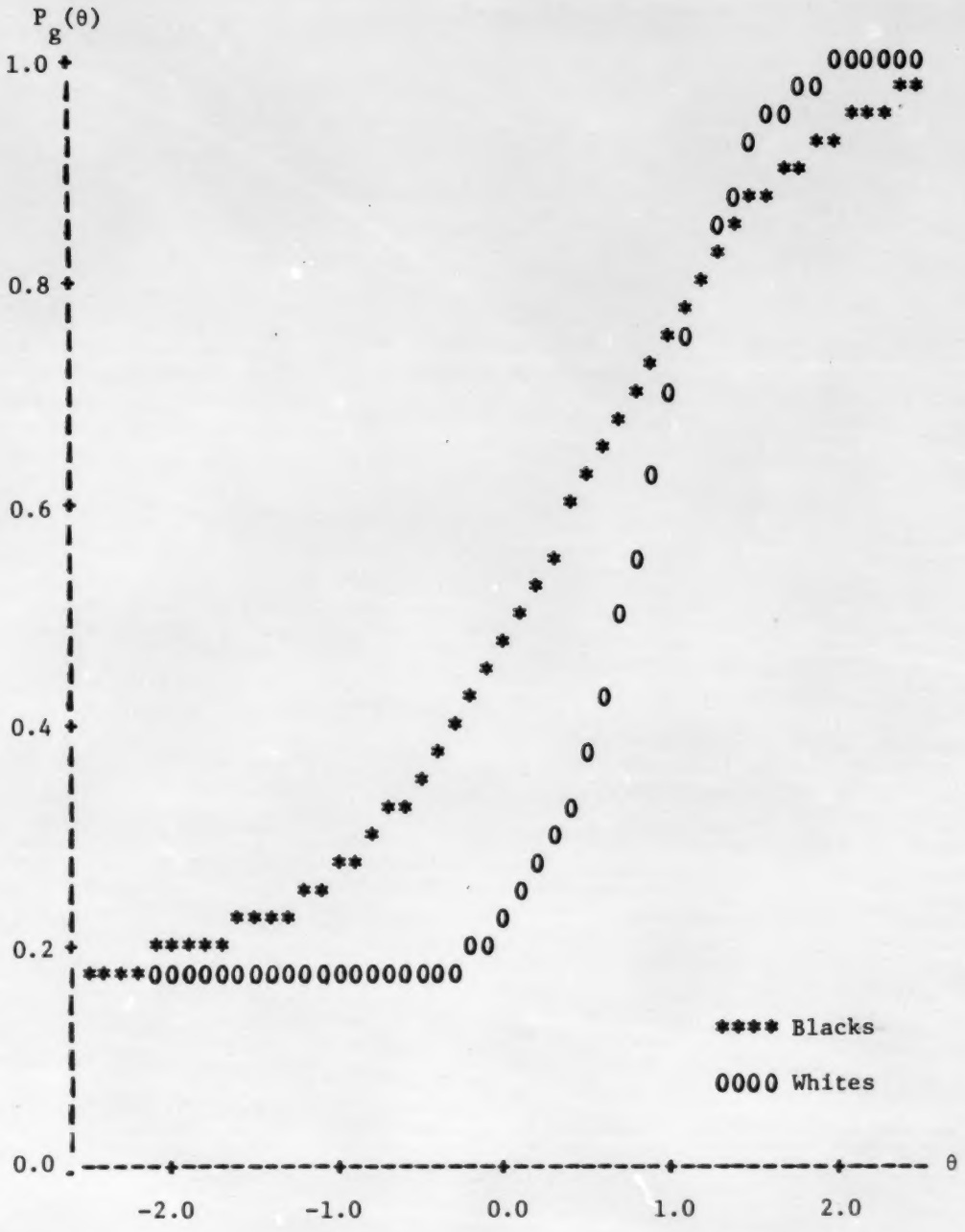

Figure 5. ICC's with Significantly Different Estimates of a_g and b_g (Item 51).

bias in favor of the members of the racial group identified. However, one might argue that this selection, by identifying the character as a black, makes the black students have as high motivation as the white students to read the selection carefully, allowing fair measurement of reading skills of both groups. If this is true, in spite of the results of the significance tests, it should be concluded that other results from the tests are not reliable either, because that argument implies bias of the whole test and falsity of equating. This again tells us how crucial to the assessment of bias is the initial assumption that the whole test is generally unbiased.

CHAPTER IV

DISCUSSION

The method proposed in this study for investigating item bias can also be used for examining the adequacy of equating based on the three-parameter logistic model. Furthermore, equating based on the one- or two-parameter logistic model can also be evaluated by the same method, since the significance test developed here can be applied when c_g is set equal to zero and when a_g is set equal to a constant value.

Both in item bias investigation and in equating, however, caution should be taken, since the distributions on which the significance tests are based are approximate distributions. It should be kept in mind that not only the approximations about the forms of the distributions and their standard deviations but also the assumptions of independence of item parameter estimates are made. In particular, the assumptions of independence involve independence of estimates of a_g, b_g, and c_g for each item and independence of estimates among different items.

The first type of independence assumption has been empirically supported: Although estimated values of a_g and b_g often have a significant positive correlation over items, this relationship is due to the nature of the θ scale, not to faulty parameter estimation procedures

(Lord, 1975a). That is, given the true values of a_g and b_g for an item, the estimates of the parameters do not exhibit any appreciable dependence.

On the other hand, the second assumption of independence would truly be satisfied only if one took a random sample of examinees for each item separately. Obviously, however, this kind of sampling invalidates the model and makes it impossible to obtain any meaningful parameter estimate.

It should also be noted that a sample of relatively large size is required to obtain meaningful results from the proposed method. One aspect of this requirement pertains to goodness of normal approximations of the distributions; that is, the approximations will be better when one takes a larger sample. Another aspect of the requirement pertains to accuracy of estimation of c_g. Given the same sample size, c_g tends to be estimated poorly compared to a_g and b_g, and the ML method occasionally produces unacceptable values of c_g when sample size is small and items are relatively easy for the examinees (Lord, 1975b).

LOGIST is equipped with a rather complicated procedure to avoid wild fluctuations of estimates of c_g. Among other things, it utilizes the approximate standard error of estimates of c_g and the value of $2/a_g - b_g$. Lord (1975b) showed that c_g is poorly estimated when $2/a_g - b_g$ is greater than two. Basically, LOGIST fixes the c_g whose standard error is large and/or for which $2/a_g - b_g \geq 2$ at the median value of c_g computed from the items for which $2/a_g - b_g < 2$. Given a certain sample size, both the standard error and the value of

$2/a_g - b_g$ can be reduced if the item difficulty level is increased. And, given a certain set of items, the standard error can be reduced by taking a larger sample size.

Regarding the sample size, the authors of LOGIST stated that "best results can be obtained with a reasonably large number of items, say 40 or more, and a reasonably large number of examinees say 1000 or more" (Wood et al., 1976, p. 5). This suggestion was made for general uses of the program. When applying the procedure proposed in the present study, it seems desirable to have a sample of two or three thousand since the sampling distributions have been derived for "genuine" ML estimators that are free from such adjustments as assigning the median value to some estimates of c_g.

There is another issue about item bias which has not been discussed so far in this paper. That is a problem of balancing effects of biased items. Based on Potthoff's (1966) suggestion, Laksana (1979) defined an index of balance in the whole test as follows:

number of items favoring one group / total number of items
- number of items favoring the other group / total number of items.

This index may be useful in the situations where an external criterion is available. However, in item bias investigation without an external criterion such as the one done by Laksana and the one proposed in this study, discussion of balance is inappropriate. That is, balance or unbiasness in the whole test is assumed in the first place. Furthermore, since such investigation focuses on deviations of item

characteristics from what is expected from the characteristics of the majority of the items in the test, the proportion of the items biased in one direction should be approximately equal to the proportion of the items biased in the other direction. Table 3 in Chapter III provides an illustration. Thirty-one out of sixty items have negative critical ratios for b_g and the other twenty-nine items have positive ratios. And, at the five percent level of significance, eight items are judged as biased against blacks and nine items against whites. This is a reasonable outcome expected from the equating procedure employed. Any other interpretation of these results would simply call into question the basic assumption that the majority of the items are unbiased.

BIBLIOGRAPHY

Anastasi, A. Psychological testing (4th ed.). New York: Macmillan, 1976.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Cramér, H. Mathematical methods of statistics. Princeton, N. J.: Princeton University Press, 1946.

Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper & Row, 1970.

Green, D. R. Reducing bias in achievement tests. Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1976. (ERIC Document Reproduction Service No. ED 126 126)

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eigner, D. R., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Hogg, R. V., & Craig, A. T. Introduction to mathematical statistics (4th ed.). New York: Macmillan, 1978.

Jensen, A. R. Test bias and content validity. Phi Delta Kappan, December 1976, 58, 340-346.

Kendall, M. G., & Stuart, A. The advanced theory of statistics (Vol. 2, 3rd ed.). London: Griffin, 1973.

Laksana, S. Application of analysis of variance approach and item characteristic curve approach for assessing item bias in the ITBS Form 7. Unpublished doctoral dissertation, University of Iowa, 1979.

Lord, F. M. Item characteristic curves estimated without knowledge of their mathematical form--A confrontation of Birnbaum's logistic model. Psychometrika, 1970, 35, 43-50.

Lord, F. M. The 'ability' scale in item characteristic curve theory. Psychometrika, 1975, 40, 205-217. (a)

Lord, F. M. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters (ETS RB 75-33). Princeton, N. J.: Educational Testing Service, 1975. (b)

Lord, F. M. A study of item bias, using item characteristic curve theory. Paper presented at the meeting of the International Association for Cross-Cultural Psychology, Tilburg, The Netherlands, July 1976. (In Y. H. Poortinga (Ed.), Basic problems in cross-cultural psychology, Amsterdam: Swets & Zeitlinger, 1977.)

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Novick, M. R., & Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw-Hill, 1974.

Pine, S. M. Applications of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), Applications of computerized adaptive testing (Research Report 77-1). University of Minnesota, Psychometric Methods Program, Department of Psychology, 1977.

Potthoff, R. F. Statistical aspects of the problem of biases in psychological tests (Institute of Statistics Mimeo Series No. 479). Chapel Hill: University of North Carolina, Department of Statistics, 1966.

Reschly, D. J., & Sabers, D. L. Analysis of test bias in four groups with the regression definition. Journal of Educational Measurement, 1979, 16, 1-9.

Rudner, L. M. An approach to biased item identification using latent trait measurement theory. Paper presented at the meeting of the American Educational Research Association, New York, April 1977. (ERIC Document Reproduction Service No. ED 137 337)

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the meeting of the American Educational Research Association, Washington, D. C., April 1975. (ERIC Document Reproduction Service No. ED 106 359)

Shiba, S. Construction of a scale for acquisition of word meanings (Bulletin of Faculty of Education, Vol. 17). University of Tokyo, 1978. (In Japanese)

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (ETS RM76-6). Princeton, N. J.: Educational Testing Service, 1976.

IOWA TESTING PROGRAMS

OCCASIONAL PAPERS

Coffman, William E. Moratorium? What Kind? (ED 099 408)

Coffman, William E., & Mathews, Walter M. Narrative Reports and the Iowa Tests of Basic Skills.

Forsyth, Robert A., & Feldt, Leonard S. On the Validity of the I.T.E.D. as an Aid in Program Evaluation. (ED 142 558)

Gulliksen, Harold O. Applications of Psychological Scaling Methods.

Lee, Lai-Min Paul, & Coffman, William E. A Study of the "I Don't Know" Response in Multiple-Choice Tests. (ED 141 371)

Novick, Melvin R. A Course in Bayesian Statistics. (out of print) (In: American Statistician, May 1975, Vol. 29, No. 2, pp. 94-98)

Hood, Joyce, Leslie, Lauren, & Kendall, Janet Ross. Instructional Effectiveness in Primary Grade Reading: A Pilot Study.

Petersen, Nancy S., & Novick, Melvin R. An Evaluation of Some Models for Test Bias. (out of print) (In: Journal of Educational Measurement, Spring 1976, Vol. 13, No. 1, pp. 3-31)

Kendall, Janet Ross, & Hood, Joyce. Instructional Practices and Easy-to-Change Surrounding Conditions Variables in Effective Primary Reading Programs. (ED 123 557)

Petersen, Nancy S. An Expected Utility Model for "Optimal" Selection. (out of print) (In: Journal of Educational Statistics, Winter 1976, Vol. 1, No. 4, pp. 333-358)

Lindley, D. V. The Effect of Ethical Design Considerations on Statistical Analysis.

Lindley, D. V. A Class of Utility Functions. (In: Annals of Statistics, Vol. 4, No. 1, January 1976)

Lindley, D. V. Probability and Medical Diagnosis; and The Role of Utility in Decision-Making.

Lindley, D. V. Inference for a Bernoulli Process: A Bayesian View; and The Future of Statistics: A Bayesian 21st Century.

Shigemasu, Kazuo. Development and Validation of a Simplified M-Group Regression Model. (In: Journal of Educational Statistics, Summer 1976, Vol. 1, No. 2, pp. 157-180)

Cantor, Gordon N. Sex and Race Effects in the Community Behavior of Upper-Elementary-School-Aged Children. (ED 115 658)

Forsyth, Robert A. Describing What Johnny Can Do.

Sedere, M. U., & Feldt, Leonard S. The Sampling Distribution of the Kristof Reliability Coefficient, the Feldt Coefficient, and Guttman's λ_2. (In: Journal of Educational Measurement, Spring 1977, Vol. 14, No. 1, pp. 53-62)

Coffman, William E. Those Achievement Tests - What For?

Hunt, David E. Teachers are Psychologists, Too: On the Applications of Psychology to Education. (Ed 129 651)

Johnson, Sylvia T. Self-Concept in a School Setting: Construct Validations by Factor Analysis. (ED 142 573)

Coffman, William E., & Shigemasu, Kazuo. Appraising School Effectiveness Using a Bayesian Method. (ED 159 220)

Coffman, William E. Classical Test Development Solutions.

Olsen, Scott A. An Investigation of Possible Test Floor Effects by Comparing Score Distributions for Students Tested with Two Different Levels of a Test Battery at Four Different Times over a Two-Year Interval.